PCI-SIG ENGINEERING CHANGE REQUEST

TITLE:	_DSM additions for Runtime Device Power Management
DATE:	May, 2017
AFFECTED DOCUMENT:	PCI Firmware Specification, Revision 3.2
SPONSOR:	Rob Gough, Intel Corporation

Part I

1. Summary of the Functional Changes

This ECN adds two capabilities by way of adding functions to the PCI Firmware Spec defined _DSM definition.

The first is a mechanism to allow for a PCIe device driver to request/negotiate with the platform to provide aux current (e.g. on the 3.3Vaux pin of a CEM slot) while the device is in $D3_{cold}$, or to prevent core power removal in the $D3_{cold}$/S0 state.

The second is a mechanism to convey from a PCIe device driver to the platform to insert a delay between the receipt of the PME_TO_Ack message at an upstream port, and the subsequent assertion of PERST#, during the transition to $D3_{cold}$, during a runtime D3 transition. It does not apply during transitions from the ACPI operational state to an ACPI sleeping state (S3/S4/S5).

2. Benefits as a Result of the Changes

The added function definitions to the _DSM allow platform and card designers to overcome some existing limitations, allowing for greater flexibility for devices desiring to support runtime $D3_{cold}$.

Some devices may require more than the maximum form-factor-specific auxillary current (375mA currently permitted by the CEM 3.0 specification). Some devices may not be able to support power removal in $D3_{cold}$, but still operate at a lower power level than when in $D3_{hot}$.

Some devices may require a delay between the time they send the PME_TO_Ack message (preparing for link L2/L3 entry) and the assertion of PERST#.

Adding support for either feature permits support for runtime device power management on a wider set of devices, reducing the amount of power consumed by a system in the ACPI operational state.

3. Assessment of the Impact

The added functions are applicable to the PCIe Bus Driver and the ACPI interface (platform firmware), and potentially board power delivery/voltage regulator designs, if implemented.

Devices/cards installed in platforms that do not support one or both of these _DSM methods may choose not to implement support for $D3_{cold}$ during an ACPI operational state.

4. Analysis of the Hardware Implications

There is no required impact to device or platform HW to support this ECN. This is an optional capability.

If the aux current limitation is increased by implementation of the first _DSM function described in this ECR, board power delivery solutions may need to be updated to support additional power requirements requested by software.

5. Analysis of the Software Implications

No changes are required, as the features described in this ECR are optional.

If either or both of these features are required by a device, the device driver for that device will need to be updated to support invoking the corresponding _DSM function within the scope of that device in the ACPI namespace.

If either or both of these features are supported on a host platform, system firmware must be modified to add the desired _DSM functions to the existing PCI _DSM Control Method.

If the auxillary current increase function is supported, system firmware (ACPI BIOS) must manage/budget the current based on capabilities of the platform, on a per-slot basis.

If the PERST# delay function is implemented, system firmware (ACPI BIOS) must ensure that the requested delay is observed as part of the $D3_{cold}$ entry sequence already implemented.

6. Analysis of the C&I Test Implications

This is an optional feature, not required. There are no prescribed HW registers required to enable/control this feature.

Part II

Detailed Description of the change

Update table 4-7 in section 4.6 as follows:

4.6 _DSM Definitions for PCI

_DSM (Device Specific Method) is defined in the ACPI 3.0 (or later) Specification. This object is a control method that enables devices to provide device specific control functions that are consumed by the device driver. Table 4-7 below lists the UUID, revision, and function definitions.

Table 4-7: _DSM Definitions for PCI

UUID	Revision	Function	Description
E5C937D0-3553-4d7a-9117-EA4D19C3434D	…	…	…
	4	0Ah	Request $D3_{cold}$ Aux Current Limit
	4	0Bh	Add PERST# Assertion Delay

Add section 4.6.x as follows:

4.6.x Request $D3_{cold}$ Aux Current Limit

This function describes how a device driver can convey its current requirements $D3_{cold}$ to the host platform when the device is in $D3_{cold}$. The system firmware responds with a value indicating whether the request can be supported. The current request is specific to the Auxillary power supply; core power may be removed while in $D3_{cold}$. A device must not draw any more current than what has been negotiated via this mechanismafter entering $D3_{cold}$.

A device driver may invoke this function multiple times, either to determine the maximum available current, to retry a request that was temporarily rejected, or to modify (raise or lower) the amount of required current.

For a Multi-Function Device, the driver for Function 0 is required to report an aggregate current requirement covering all functions contained within the device.

Earlier current limits are superseded when this function grants a request. In all other cases, this function has no effect on earlier current limits. This includes grants from earlier calls of this function as well as requirements for the D3 PM state in the form factor and PCI Express Base specifications.

Location:

Support for this function is only applicable for a _DSM Control Method within the scope of a PCI Express Downstream Port.

For bus hierarchies where multiple Functions reside beneath the PCI Express Downstream Port supporting this _DSM function, system software is responsible for tracking and aggregating requests from child devices and requesting the sum of the requested current limits.

Arguments:

Arg0: UUID: E5C937D0-3553-4d7a-9117-EA4D19C3434D

Arg1: Revision ID: 4

Arg2: Function Index: 0Ah

Arg3: Integer. The value of the integer is the amount of current requested, in mA, for the auxillary power supply. The minimum allowed value is 0; the maximum allowed value varies by form factor (e.g., for the CEM 3.3vAux, the maximum value is 1000, indicating 1 A). A value of 8000 0000h signifies that the hierarchy connected via the slot cannot support core power removal when in $D3_{cold}$ while the system is in S0.

Return:

Integer:

0h - Denied. Indicates that the platform cannot support the current requested. Software may retry the request with either a lower current request, or require no power removal. This is also the value returned to signify any error with the request.

1h - Granted. Indicates that the device is permitted to draw the requested auxillary current.

2h - No core power removal. Indicates that the platform will not remove core power from the slot while the system is in S0. This value may be returned even though it was not requested that core power remain on, based on requirements of other devices within the platform or other platform configuration variations. This is the only valid return value when Arg3 is 8000 0000h.

3h to 10h Reserved.

11h to 1Fh - Retry, with interval. Indicates that the platform cannot support the current requested at this time, but that it may be able to in the future. Software should retry the request in the number of seconds corresponding with the lower 4 bits of the return value (1 to 15). Firmware is not permitted to return a value in this range more than once for each _DSM instance (located within the ACPI Namespace of a single downstream port DeviceObject), unless there is a subsequent invocation of this function before the previously returned retry interval has expired.

All other values are Reserved.

IMPLEMENTATION NOTE
Platform Firmware Budgeting of Aux Current Availability
Platform firmware must not grant more current than what is available within the system.

For example, a board may be designed with 4 CEM slots (one x16 slot, one x4 slot, and two x1 slots). The board may implement a power delivery circuit capable of supplying 2 A of current for the 3.3Vaux rail supplying all 4 slots. The 3.3Vaux pins on each CEM slot can supply 1 A each.

Platform firmware may use the retry mechanism to prioritize requests from devices in preferred slots in the following manner:

- Requests from a device in the highest priority slot (e.g., x16) are granted immediately, if available.
- Requests from devices in lower priority slots (e.g., x2, x1) are initially rejected, with a retry interval inversely proportional to the slot priority. For instance, if the x2 slot is higher priority than the x1 slots, so the retry interval for the x2 slot may be 4 seconds, and the x1 slots may be 8 and 10 seconds.
- As requests are granted, the granted values are subtracted from the available budget.
- Retried requests are granted based on the remaining current budget, or denied if insufficient current budget is available. Another retry is not permitted.
- When there insufficient current budget for a request, firmware may choose to keep core power on and return no power removal (2h).

Add section 4.6.y as follows:

4.6.y Add PERST# Assertion Delay

This function is used to convey the requirement for a fixed delay in timing between the time the PME_TO_Ack message is received at the PCI Express Downstream Port that originated the PME_Turn_Off message, and the time the platform asserts PERST# to the slot during the corresponding Endpoint's or PCI Express Upstream Port's transition to $D3_{cold}$ while the system is in an ACPI operational state. This delay is not guaranteed to be applied during the transition to a system sleeping state. Assertion of PERST# may be an indicator that power to the device is about to be removed. There is no guaranteed delay between assertion of PERST# and power removal. If any Function on the device is armed for wake, auxiliary power rails required to supply wake logic will not be turned off.

Host platforms implementing this feature must ensure that the delay is observed by the device that is being transitioned to $D3_{cold}$. Once set, this delay will be met on every $D3_{cold}$ transition of the device. This function may be invoked multiple times, allowing the delay value to be changed at any time, so that the new value can be applied for the next $D3_{cold}$ transition of the device.

Location:

Support for this function is only applicable for a _DSM Control Method within the scope of a PCI Express Downstream Port terminated by any sort of add-in slot/connector.

For bus hierarchies where multiple Functions reside beneath the PCI Express Downstream Port supporting this _DSM function, system software is responsible for tracking and

aggregating requests from child devices and requesting the maximum of the requested delay values.

Arguments:

Arg0: UUID: E5C937D0-3553-4d7a-9117-EA4D19C3434D

Arg1: Revision ID: 4

Arg2: Function Index: 0Bh

Arg3: Integer. The value, in microseconds, of the delay needed between the PME_TO_Ack message receipt at the PCI Express Downstream Port that originated the PME_Turn_Off message, and the subsequent assertion of PERST# on the corresponding slot. This delay is injected during the transition to during a runtime D3 transition. It does not apply during transitions from the ACPI operational state to an ACPI sleeping state (S3/S4/S5). The maximum permitted requested delay is 10 ms.

Return:

Integer: Current delay value, in microseconds. If the returned value is not identical to the requested value, this signifies an error.